Affix Learning
Zulu and German

Uwe Quasthoff

Universität Leipzig
Institut für Informatik
quasthoff@informatik.uni-leipzig.de

Learning of prefixes and suffixes with unknown stems

Quality of machine learning methods for this task: Which classifier works well, and how much training data is needed for a certain quality?

Available training data for Zulu (limited):

- Set 1: isiZulu NCHLT Annotated Text Corpora (Language Resource Management Agency, 2013) -lemmatised, POS tagged and morphologically decomposed corpora. 19,439 words.
- Set 2: Ukwabelana word list - segmentations and labelled morphological analyses described by Spiegler et al. (2010). 9,224 words.

Complex morphology of Zulu

Class-dependent prefixes: based on 18 noun classes, express grammatical correlation in verbs (subject and object prefixes), adjectives, etc.

Class-independent suffixes: not based on noun classes, therefore fewer in number and variation.

- Nouns > diminutive, augmentative, locative morphemes;
- Verbs > verb terminative (pos/neg, tense) and extension morphemes.

izindlwana	**ezikude**	**azibonakali**	**ngamehlo**	**endleleni**
izin-**dlu**-ana	ezi-**kude**	azi-**bon**-akal-i	nga-ama-**ihlo**	e-in-**dlela**-ini
houses-small	that-are-far	they-not-visible	with-eyes	from-road
Small houses that are far away are not visible with the eyes from the road				

Classification example with TiMBL

TiMBL (Tilburg Memory-Based Learner) is an open source classifier: decision-tree-based k-nearest neighbor classification.

Attributes chosen for the morphology task:

- Character n-grams at word beginnings (n=1, …, 7)
- Character n-grams at word endings (n=1, …, 7)
- Uppercase beginning?
- POS tag (if available)

word	up	l1	l2	l3	l4	l5	l6	l7	r7	r6	r5	r4	r3	r2	r1	POS	pre	suf
abahlali	0	a	ab	aba	abah	abahl	abahla	abahlal	bahlali	ahlali	hlali	lali	ali	li	i	V	aba	i
abakwenziwe	0	a	ab	aba	abak	abakw	abakwe	abakwen	wenziwe	enziwe	nziwe	ziwe	iwe	we	e	V	abakw	iwe
esesemabangeni	0	e	es	ese	eses	esese	esesem	esesema	bangeni	angeni	ngeni	geni	eni	ni	i	N	esesema	eni
lainvume	0	l	la	lai	lain	lainv	lainvu	lainvum	ainvume	invume	nvume	vume	ume	me	e	N	lain	*
ngaukugcwalisa	0	n	ng	nga	ngau	ngauk	ngauku	ngaukug	cwalisa	walisa	alisa	lisa	isa	sa	a	V	ngauku	isa
ngauMsombuluko	1	n	ng	nga	ngau	ngauM	ngauMs	ngauMso	mbuluko	buluko	uluko	luko	uko	ko	o	N	ngau	*
saukufakela	0	s	sa	sau	sauk	sauku	saukuf	saukufa	ufakela	fakela	akela	kela	ela	la	a	V	sauku	ela
ukubuya	0	u	uk	uku	ukub	ukubu	ukubuy	ukubuya	ukubuya	kubuya	ubuya	buya	uya	ya	a	V	uku	a

Quality of Results

With about 8,000 training words we achieve (on average)

- 94% accuracy for suffix detection
- 75% accuracy for prefix detection
- 90% accuracy for POSprefix detection

Task	Entropy	Percentage used as training data								
		10	20	30	40	50	60	70	80	90
Number of training data		923	1,846	2,769	3,691	4,613	5,535	6,457	7,379	8,301
RMA-prefix	7.08	67	70	71	72	73	73	74	76	78
RMA-POS	1.00	84	87	89	89	90	90	91	92	92
RMA-suffix	2.10	91	93	94	94	95	96	96	96	96
Spiegler-prefix	5.93	61	63	63	63	64	64	66	67	72
Spiegler-POS	0.85	80	84	84	85	86	87	87	87	87
Spiegler-suffix	2.54	85	88	90	91	91	91	92	92	92

Error Analysis

- Stems with a certain structure,
- ambiguities,
- foreign stems,
- inconsistencies in training data

Zulu example

Error	Examples	Explanation
engiyi / engi	engiyizwayo	general difficulty to recognize monosyllabic verb roots
sengiyam / sengiya	sengiyam-esaba	general difficulty to recognize verb roots starting with a vowel
baka / ba	bakageorge	difficulty to recognize foreign noun stems

German example

Error	Examples	Explanation
en/n	Reisen, Terminen	general difficulty to recognize nouns ending in –e
n/en	Pannen, Ruinen, Vorlieben	general difficulty to recognize nouns ending in –e
-/s	Orleans, Glas	short or foreign words

Application to other languages: Generalizations

Some problems are typical for a language or language family. For Zulu (and the contrast language German) we found:

- Suffix detection is simpler than prefix detection.
- If the classification task is simple (due to morphological regularities), 1,000-5,000 training words can imply an accuracy of >90%
- Approach will work for a large number of languages
- Other tasks on similar data: lemmatization, hyphenation, decomposition of compounds, subject area, pronunciation